Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Desenvolvimento de metodologias para avaliação do resultado de campanhas de Marketing Direto a clientes no setor das telecomunicações

Versão Pública

Mariana Alexandra Moreira Henriques

Mestrado em Estatística

2011/2012

Universidade de Lisboa

Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Desenvolvimento de metodologias para avaliação do resultado de campanhas de Marketing Direto a clientes no setor das telecomunicações

Versão Pública

Mariana Alexandra Moreira Henriques

Dissertação orientada pelo

Prof. Dr. João Gomes

Mestrado em Estatística

2011/2012

Agradecimentos

Agradecimento especial à empresa onde realizei o estágio pela excelente oportunidade que me concederam e a todos os colaboradores por todos os conselhos e ajuda que iram certamente ser me úteis no futuro, em particular à orientadora.

Gostava ainda de agradecer ao meu orientador João Gomes pela paciência e o conhecimento e pelo tempo gasto a responder às minhas questões e dúvidas.

A todos os professores que passaram ao longo do meu percurso académico e incentivaram o meu gosto pela Estatística.

Não podia deixar de agradecer a todos os meus colegas e amigos, pela força e o incentivo que me transmitiram ao longo destes anos. Em especial, ao meu namorado que sempre acreditou em mim e me motivou ao máximo.

Finalmente aos meus pais e irmãs que sempre me deram tudo e sem eles não conseguiria chegar ao fim.

Resumo

O papel das tecnologias tem vido a aumentar no nosso dia-a-dia, com um rápido desenvolvimento a cada momento que passa, não sendo o ramo das telecomunicações alheio a esta constante mudança, a este dinamismo. Por isso, é preciso atender às múltiplas necessidades da sociedade e ter a capacidade de apresentar aos clientes as melhores propostas tendo em conta os seus desejos e necessidades.

As tecnologias atuais proporcionam a produção de grande volume de dados, a maioria com um número muito elevado de variáveis. É frequente utilizarem-se técnicas de prospeção de dados (*data mining*) e de aprendizagem automática (*machine learning*). Estas técnicas exigem tarefas de otimização que implicam a perícia do utilizador na escolha desses critérios.

De forma a desenvolver modelos para cada campanha, é importante conseguir prever como os clientes se comportam, as suas particularidades, e, consoante estas características, determinar aquela (s) que o cliente estará mais propenso a aceitar. Estas, irão ser propostas ao cliente consoante o seu perfil. Com a avaliação dos resultados obtidos pretende-se detetar as campanhas que fazem sentido continuar no mercado e aquelas que já não têm utilidade, assim como introduzir novas campanhas.

Para os objetivos expostos no parágrafo anterior foram utilizadas algumas metodologias, como por exemplo: introdução de novas variáveis, e (ou) eliminação de outras nos modelos adaptativos, aplicação de novas técnicas de análise como a regressão logística ou a criação de grupos de controlo.

Palavras-Chave: Telecomunicações, Modelos Lineares Generalizados, Modelos Adaptativos, Regressão Logística, Grupos de Controlo

Abstract

The role of technology is increasing in our daily lives and developing fast, at each moment. The telecommunications world is not a strange to this shifting change, to this dynamics. Due to this, it is necessary to attend to the multiple needs of society and be able to have the capacity to present the best proposals according to each individual wishes and needs.

The present technologies lead to the production of large data bases, most of which with a great number of variables. In face of these it is usual to use tools as data mining and machine learning. These techniques require optimization tasks which imply the user's skill in the choice of these criteria.

In order to develop the models for each campaign, it is important to have the ability to predict how the clients behave, their particularities and, according to these characteristics, to determine which ones the client will be more prone to accept. These will be proposed to the client depending on his profile. With the evaluation of the obtained results we pretend to find the campaigns which make more sense to stay in the market and those that no longer have use, as well as to introduce new campaigns.

For the objectives exposed in the previous paragraph, some methodologies were used such as: introduction of new variables and (or) removal of others in the adaptive models, use of new analysis techniques like logistic regression or development of control groups.

Keywords: Telecommunications, Generalized Linear Models, Adaptive Models, Logistic Regression, Controls Groups

Índice

Índice de Tabelas

Índice de Figuras

1. Enquadramento

1.1 Introdução

Num mundo em constante mudança, onde tudo se move a um ritmo frenético e onde a novidade do hoje é o obsoleto do amanhã, as telecomunicações fazem parte deste grupo. Num século onde estas desempenham cada vez um papel mais importante neste fluxo de mudanças e inovações, onde a sociedade interage cada vez mais e exige comunicações rápidas a um custo cada vez menor, é essencial atender a estas necessidades e conseguir disponibilizar soluções a um ritmo rápido e fácil para o cliente.

Por tudo isto, é fundamental estudar e avaliar as campanhas de forma a poder propor melhorias, decidir quais e que fazem mais sentido continuar e aquelas que não tem utilidade. Desenvolver ferramentas que otimizem os modelos inerentes a algumas das campanhas, avaliando assim o perfil do cliente de modo a atender às necessidades e desejos destes.

1.2 Obtenção dos Dados

Ao longo do projeto, foram utilizados vários dados distintos de onde resultaram três estudos: a criação de um modelo para uma determinada campanha, a análise das campanhas ao longo de um intervalo de tempo de 3 meses e a análise comparativa das taxas de sucesso para um grupo de campanhas.

Para a replicação do modelo de uma campanha específica de clientes foram obtidas informações acerca destes entre fevereiro e abril de 2012, cedidos gentilmente pela empresa onde o estágio foi realizado. Desta amostra foram retiradas as variáveis que incluíam características sobre o telefone, nomeadamente, a tecnologia, o tipo de geração, a antiguidade, entre outras, bem como, o tarifário de cada cliente, o montante de cada carregamento e a faturação de chamadas.

Na construção do modelo utilizou-se o *software R*.

Para a análise realizada a todas as campanhas compreendidas no período entre os meses de junho de 2011 a setembro de 2011 foi considerada a base total de clientes e extraída informação como o tipo de tarifário de cada cliente, o histórico de campanhas a que cada

cliente foi alocado e os canais de comunicação em que foi feito o contacto e a frequência de contacto.

Desta abordagem resultou um elevado número de dados e variáveis. Perante este cenário, foi essencial a utilização de um programa de base de dados, o *Oracle SQL Developer*, que permitiu chegar a uma sólida base de dados onde este estudo foi exequível.

Numa outra abordagem, analisou-se os modelos adaptativos para um determinado grupo de campanhas com vista a otimizar os modelos. Como cada modelo tem como propósito estimar a propensão do cliente a aceitar uma campanha, foi feito ainda um estudo à fórmula de priorização de forma a estabelecer a ordem pela qual as campanhas são transmitidas ao cliente.

Na análise comparativa das taxas de sucesso para um grupo de campanhas utilizaram-se dados referentes aos meses entre abril de 2011 e março de 2012. O seu objetivo é comparar as taxas de sucesso entre os dois canais de contacto, *call center* e lojas.

Por fim, outra análise feita foi a criação de metodologias de grupos de controlo. O objetivo aqui foi avaliar a eficiência das campanhas entre si.

2. Metodologias

2.1 Conceitos Introdutórios

Neste capítulo vão ser apresentados os conceitos necessários ao longo de todo o projeto.

2.1.1 Teorema da Probabilidade Total

O teorema da Probabilidade Total, pode ser descrito como sendo *S* um acontecimento e $\{D_n\}_{n\in\aleph}$ uma partição do universo Ω em acontecimentos, então:

$$P(S) = \sum_{n\in\aleph} P(S|D_n)P(D_n)$$

(Pestana & Velosa, 2008)

2.1.2 Teorema de Bayes

O Teorema de Bayes não é mais do que um corolário do Teorema da Probabilidade Total, onde $\{D_n\}_{n\in\aleph}$ continua a ser uma partição do universo Ω em acontecimentos, e *S* um acontecimento de Ω.. Então:

$$P(D_i|S) = \frac{P(S|D_i)P(D_i)}{\sum_{n\in\aleph} P(S|D_n)P(D_n)}$$

(Pestana & Velosa, 2008)

2.1.3 Modelo Naive Bayes

O método Naive Bayes é um algoritmo baseado no Teorema de Bayes cujo objetivo é obter um classificador que produza valores de distribuição de probabilidade para possíveis valores de *Y* em cada instante de *X*. O modelo assume que as variáveis aleatórias $\boldsymbol{X_1, X_2, \dots, X_n}$ são condicionalmente independentes entre si dado a variável aleatória *Y*. Esta premissa simplifica a representação de $\boldsymbol{P(X|Y)}$ e consequentemente a estimação.

Ao considerarmos assim uma variável aleatória *Y* e *X* um vetor que contem *n* atributos, ou seja, $\boldsymbol{X} = \langle \boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n} \rangle$ onde a variável aleatória $\boldsymbol{X_i}$ é o i-ésimo atributo de *X*, podemos aplicar a definição de probabilidade condicional para $\boldsymbol{P(X|Y)}$:

$$\boldsymbol{P(X_1 \dots X_n|Y) = P(X_1\,|Y) \dots P(X_n|Y) = \prod_{i=1}^{n} P(X_i\,|Y)}$$

Aplicando o Teorema de Bayes $\boldsymbol{P(Y = y_i|X_1 \dots X_n)}$ é representada como:

$$\boldsymbol{P(Y = y_i|X_1 \dots X_n) = \frac{P(X_1 \dots X_n|Y = y_i)P(Y = y_i)}{\sum_j P(X_1 \dots X_n|Y = y_j)P(Y = y_j)}}$$

Como as variáveis aleatórias $\boldsymbol{X_1, X_2, \dots, X_n}$ se supõem condicionalmente independentes entre si dado a variável aleatória *Y*, podemos rescrever a equação como:

$$\boldsymbol{P(Y = y_i|X_1 \dots X_n) = \frac{\prod_{i=1}^{n} P(X_i\,|Y = y_i)\, P(Y = y_i)}{\sum_j P(Y = y_j) \prod_{i=1}^{n} P(X_i\,|Y = y_j)}}$$

Esta equação é fundamental para o método Naive Bayes pois mostra como calcular a probabilidade para qualquer valor de *Y* dado os valores observados de $\boldsymbol{X} = \langle \boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_n} \rangle$ e as distribuições de probabilidade de *Y* e de $X_i|Y$ estimadas dos dados.

O método Naive Bayes estima diretamente *Y* e $X|Y$, ao contrário da Regressão Logística em que estimamos a distribuição de $Y|X$, diretamente descriminada para valor target *Y* dado um instante *X*. (Mitchell, 1997)

2.1.4 **Família Exponencial**

De acordo com a literatura (Turkman & Silva, 2000) diz-se que uma variável aleatória *Y* tem distribuição pertencente à família exponencial (θ,ϕ) se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever da seguinte forma:

$$f(y|\theta, \phi) = \; exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

onde θ e ϕ são parâmetros escalares e *a(.)*, *b(.)* e *c(.,.)* são funções reais conhecidas.

Na definição anterior, θ é a forma canónica do parâmetro de localização e ϕ um parâmetro de dispersão suposto, em geral, conhecido.

Prova-se que o valor médio de *Y* é dado por $E(Y) = b'(\theta)$ e a variância por $Var(Y) = a(\phi)b''(\theta)$. A variância de *Y* é o produto de duas funções, $b''(\theta)$, que depende apenas do parâmetro canónico θ (e portanto do valor médio μ), a que se dá o nome de função de variância e que se costuma representar por $V(\mu)$ e outra, $a(\phi)$, que depende apenas do parâmetro de dispersão ϕ. Há situações em que a função $a(\phi)$ toma a forma $a(\phi) = \frac{\phi}{\omega}$, onde ω é uma constante conhecida, obtendo-se portanto a variância de *Y* como o produto do parâmetro de dispersão por uma função apenas do valor médio.

Neste caso a função definida anteriormente escreve-se na forma

$$f(y|\theta, \phi, \omega) = exp\left\{\frac{\omega}{\phi}\big(y\theta - b(\theta)\big) + c(y, \phi, \omega)\right\}.$$

Existe uma grande variedade de distribuições de probabilidade conhecidas que pertencem à família exponencial onde se incluem as distribuições normais, binomiais e Poisson, entre outras (Turkman & Silva, 2000).

Exemplo: Distribuição Binomial

Dentro da família exponencial, iremos utilizar, principalmente, a distribuição binomial. Assim, se *Y* tiver uma distribuição binomial com parâmetros *m* e π, $Y \sim B(m, \pi)$, a sua função massa de probabilidade é dada por:

$$\begin{aligned} f(y|\pi) &= \binom{m}{y} \pi^y \, (1-\pi)^{m-y} \\ &= exp\left\{ln\binom{m}{y} + \, y \, ln\, \pi + (m-y)\, ln(1-\pi)\right\} \\ &= exp\left\{ ln\binom{m}{y} + y \, ln\left(\frac{\pi}{1-\pi}\right) - m \, ln(1-\pi)\right\} \end{aligned}$$

com $y \in \{0,1,\dots,n\}$.

De acordo com a notação anterior definida para a função exponencial, podemos deduzir as seguintes expressões:

- $\theta = ln(\frac{\pi}{1-\pi}) \Leftrightarrow \pi = \frac{e^\theta}{1+e^\theta}$
- $b(\theta) = -m \, ln(1-\pi)$ que se pode escrever ainda como $b(\theta) = m \, ln(1+e^\theta)$
- $c(y, \phi) = ln\binom{m}{y}$

- $b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \pi$
- $b''(\theta) = V(\mu) = \frac{e^{\theta}}{(1+e^{\theta})^2} = \pi(1 - \pi$
- $a(\phi) = 1 \Rightarrow \phi = \omega = 1$

Do exposto anteriormente, rapidamente se chega à conclusão que $E(Y) = b'(\theta) = \pi$ e $Var(Y) = b''(\theta)a(\phi) = \pi(1 - \pi)$.
(Turkman & Silva, 2000)

2.2 Modelos Lineares Generalizados

A aplicação dos modelos lineares generalizados (MLG) tem-se verificado em diferentes áreas científicas. Estes modelos permitem, por um lado, que a variável resposta não tenha distribuição normal e, por outro, que estejam relacionados com a variável resposta através de uma função de ligação.

Consideremos então, uma variável aleatória Y de interesse, que designamos por variável resposta ou variável dependente, e um vetor $x = (x_1,...,x_k)^T$ de k variáveis explicativas, também denominadas por covariáveis ou variáveis independentes. Tanto a variável resposta como as covariáveis podem ser de natureza contínua ou discreta. Os modelos lineares generalizados pressupõem que a variável resposta tenha uma distribuição pertencente a uma família particular, a família exponencial.

Os MLG são caracterizados pela seguinte estrutura:

1. Componente aleatória

Dado o vetor de covariáveis x_i, as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial, com $E(Y_i|x_i) = \mu_i = b'(\theta_i)$ para $i=1,..,n$ e, possivelmente, um parâmetro de dispersão φ não dependente de i.

2. Componente estrutural ou sistemática

O valor esperado μ_i está relacionado com o preditor linear $\eta_i = z_i^T\beta$ através da relação:

$$\mu_i = h(z_i^T \beta) \Rightarrow \mu_i = h(\eta_i)$$,ou, invertendo, $\eta_i = h^{-1}(\mu_i) = g(\mu_i)$

onde

- h é uma função monótona e diferenciável;
- $g = h^{-1}$ é denominada por função de ligação;
- β é um vetor de parâmetros de dimensão p;
- z_i é um vetor de especificação de dimensão p, função do vetor de covariáveis x_i.

Quando se verifica $\eta_i = \theta_i$ a função de ligação designa-se por ligação canónica (o parâmetro canónico coincide com o preditor linear).

Em geral, tem-se que $z_i = (1, x_{i1}, \dots x_{ik})^T$ com $k = p - 1$. Contudo, quando existem covariáveis qualitativas elas têm de ser, em geral, convenientemente codificadas à custa de variáveis binárias mudas (*dummy*); por exemplo, se uma variável qualitativa (ou fator) tem q categorias, são necessárias $q - 1$ variáveis binárias para a representar. Essas variáveis têm então de ser incluídas no vetor z. (Turkman & Silva, 2000)

A escolha da função de ligação depende do tipo de resposta e do estudo particular que se está a fazer.

Existem três etapas fundamentais para modelação através do Modelo Linear Generalizado: formulação, ajustamento e seleção e validação dos modelos.

2.3 Modelo de Regressão Logística

O Modelo de Regressão Logística é um dos casos particulares dos Modelos Lineares Generalizados.

Uma vez que a variável resposta no modelo de regressão logística é binária ou dicotómica, as respostas podem ser por exemplo: morto ou vivo, presente ou ausente, sucesso ou insucesso, em termos genéricos. Nestes modelos específicos, a componente aleatória tem distribuição binomial, as covariáveis são mistas e a função de ligação é a *logit*. (Hosmer & Lemeshow, 2000) A função *logit* é definida para um valor p entre *0* e *1* como $logit(p) = log\left(\frac{p}{1-p}\right)$.

Suponhamos que estamos então perante n variáveis resposta independentes Y_i, cuja distribuição é binomial com parâmetros *1* e π_i e escrevemos $Y_i \sim B_i(1, \pi_i)$. Uma vez mais a função massa de probabilidade é dada por:

$$f(y_i|\pi_i) = \pi_i{}^{y_i}(1-\pi_i)^{1-y_i}, \qquad \text{para } y_i = 0,1$$

a cada indivíduo *i* ou unidade experimental, está associado um vetor de especificação z_i, resultante do vetor de covariáveis x_i, $i = 1, \ldots, n$.

Como a distribuição binomial pertence à família exponencial, foi visto anteriormente que $E(Yi) = \pi_i$ e que θ_i é dado por $\theta_i = ln(\frac{\pi_i}{1-\pi_i})$.

Portanto θ_i e η_i são ambas funções lineares de π_i, e ao fazer $\theta_i = \eta_i$ sem dificuldade se chega a que:

$$\theta_i = \eta_i = z_i^T\beta \Rightarrow z_i^T\beta = ln(\frac{\pi_i}{1-\pi_i}) = logit(\pi_i)$$

conclui-se assim que a função de ligação canónica é a função *logit*.

A probabilidade de sucesso, ou seja $P(Y_i = 1) = \pi_i$, está relacionada com o vetor z_i através de $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \Rightarrow \pi_i = \frac{exp\,(z_i^T\beta)}{1+exp\,(z_i^T\beta)}$.

Prova-se que a função distribuição logística dada por $F(x) = \frac{exp(x)}{1+exp(x)}$ é uma função de distribuição com $F: \mathbb{R} \to [0,1]$. Assim, o Modelo Linear Generalizado definido pelo modelo binomial, com função de ligação canónica *logit*, é conhecido por modelo de regressão logística (Turkman & Silva, 2000, pp. 17-18). Existem outras funções de ligação como a função *probit*, *complementar log-log* e a *log-log*.

2.3.1 **Estimação dos parâmetros**

Os estimadores de máxima verosimilhança (EMV) são obtidos como solução das equações de máxima verosimilhança e é o método por norma utilizado para a estimação dos parâmetros de um modelo linear generalizado. (Turkman & Silva, 2000)

A função de verosimilhança *L(θ)* representa a distribuição conjunta dos dados observados. Uma vez encontrada a função para um determinado conjunto de dados, o método da máxima verosimilhança determina a estimação de um conjunto de parâmetros desconhecidos que maximizam a função de verosimilhança *L(θ)*.

Em geral, maximizar a função *L(θ)* é equivalente a maximizar o logaritmo de *L(θ),*que é, em termos de cálculo, mais simples. As componentes de *θ* são encontradas como solução

das equações de derivadas parciais do logaritmo da função de verosimilhança igualadas a zero, em relação à variável θ_j, j o parâmetro individual. (Kleinbaum & Klein, 2010)

Consideremos o modelo linear generalizado definido anteriormente através da função exponencial, $f(y|\theta,\phi,\omega) = exp\left\{\frac{\omega}{\phi}\left(y\theta - b(\theta)\right) + c(y,\phi,\omega)\right\}$.

A função de verosimilhança é dada como função do parâmetro β por:

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\theta_i,\phi,\omega_i) =$$

$$= \prod_{i=1}^{n} exp\left\{\frac{\omega_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i,\phi,\omega_i)\right\}$$

$$= exp\left\{\frac{1}{\phi}\sum_{i=1}^{n}\omega_i\left(y_i\theta_i - b(\theta_i)\right) + \sum_{i=1}^{n} c(y_i,\phi,\omega_i)\right\}$$

e portanto o logaritmo da função de verosimilhança (*log-verosimilhança*) é dado por

$$ln\, L(\beta) = \ell(\beta) = \sum_{i=1}^{n}\left[\frac{\omega_i\left(y_i\theta_i - b(\theta_i)\right)}{\phi} + c(y_i,\phi,\omega_i)\right]$$

$$= \sum_{i=1}^{n}\ell_i(\beta)$$

onde

$$\ell_i(\beta) = \frac{\omega_i\left(y_i\theta_i - b(\theta_i)\right)}{\phi} + c(y_i,\phi,\omega_i)$$

é a contribuição de cada observação y_i para a verosimilhança.

Admitindo que se verificam certas condições de regularidade, os estimadores de máxima verosimilhança para β são obtidos como solução do sistema de equações de verosimilhança.

$$\frac{\partial\ell(\beta)}{\partial\beta_j} = \sum_{i=1}^{n}\frac{\partial\ell_i(\beta)}{\partial\beta_j} = 0, \quad j = 1,\dots,p.$$

(Turkman & Silva, 2000)

No caso que se tem vindo a analisar em que $Y \sim B(m,\pi)$, vimos que a função massa de probabilidade é dada por:

$$f(y|\pi) = \binom{m}{y}\pi^{y}\,(1-\pi)^{m-y} = exp\left\{ ln\binom{m}{y} + y\,ln\left(\frac{\pi}{1-\pi}\right) - m\,ln(1-\pi)\right\}$$

com $y \in \{0,1,\dots,n\}$. Portanto a função de verossimilhança é dada por:

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\pi_i) = \prod_{i=1}^{n} exp\left\{ ln\binom{m_i}{y_i} + y\,ln\left(\frac{\pi_i}{1-\pi_i}\right) - m_i\,ln(1-\pi_i)\right\}$$

$$= exp\left\{\sum_{i=1}^{n} ln\binom{m_i}{y_i} + y_i\,ln\left(\frac{\pi_i}{1-\pi_i}\right) - m_i\,ln(1-\pi_i)\right\}$$

e substituindo, $z_i^T\beta = ln(\frac{\pi_i}{1-\pi_i})$, chegamos a:

$$L(\beta) = exp\left\{\sum_{i=1}^{n} ln\binom{m_i}{y_i} + y_i z_i^T\beta - m_i\,ln\left(1 - \frac{e^{z_i^T\beta}}{1+e^{z_i^T\beta}}\right)\right\}$$

$$= exp\left\{\sum_{i=1}^{n} ln\binom{m_i}{y_i} + y_i z_i^T\beta + m_i\,ln\left(1+e^{z_i^T\beta}\right)\right\}$$

Aplicando o logaritmo à função verossimilhança temos:

$$ln\,L(\beta) \;= \ell(\beta) = \sum_{i=1}^{n}\binom{m_i}{y_i} + y_i z_i^T\beta + m_i\,ln(1+e^{z_i^T\beta})$$

Derivando a função *log-verosimilhança*, temos que os estimadores de máxima verosimilhança para β são obtidos como solução do sistema de equações:

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}\frac{\partial \ell_i(\beta)}{\partial \beta_j} = 0 \quad j = 1,\dots,p.$$

O problema destas equações é que não têm solução analítica e portanto é necessário recorrer a métodos iterativos.

O método iterativo que se vai apresentar, é baseado no método de *scores de Fisher*, também denominado por método iterativo dos mínimos quadrados.

A componente *j* da função *score* é definida como sendo

$$S_j(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} s_i(\beta) = \sum_{i=1}^{n}\frac{(y_i - \mu_i)z_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}\,, j = 1,\dots,p.$$

Para famílias regulares temos que $E\big(S(\beta)\big) = 0$ e $E\big(S^T(\beta)S(\beta)\big) = -E\left[\frac{\partial^2 \ell n(\beta)}{\partial\beta\partial\beta^T}\right]$.

A matriz de covariância da função *score,* $\Gamma(\beta) = E\left[-\frac{\partial S(\beta)}{\partial\beta}\right]$ é conhecida como matriz de informação de Fisher e prova-se que o elemento genérico de ordem *(j,k)* pode ser escrito como: $-E\left(\frac{\partial^2 \ell_i}{\partial\beta_j\partial\beta_k}\right) = \frac{z_{ij}z_{il}}{Var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$.

Na forma matricial temos que $\Gamma(\beta) = Z^T W Z$, onde *W* é a matriz diagonal de ordem *n* cujo i-ésimo elemento é $\overline{\omega} = \frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{Var(Y_i)} = \frac{\omega_i\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{\phi V(\mu_i)}$.

Seja então $\hat{\beta}^{(0)}$ uma estimativa inicial para β. O processo de *scores de Fisher* utiliza o cálculo das sucessivas iteradas através da relação:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left[\Gamma\left(\hat{\beta}^{(k)}\right)\right]^{-1} s(\hat{\beta}^{(k)})$$

onde $\Gamma(.)^{-1}$, a inversa (que se supõe existir) da matriz de informação de Fisher e *s(.)* o vetor de *scores*, são calculados para $\beta = \hat{\beta}^{(k)}$.

A expressão anterior pode ser ainda escrita na forma

$$\left[\Gamma\left(\hat{\beta}^{(k)}\right)\right]\hat{\beta}^{(k+1)} = \left[\Gamma\left(\hat{\beta}^{(k)}\right)\right]\hat{\beta}^{(k)} + s(\hat{\beta}^{(k)})$$

O lado direito desta equação acima é um vetor com elemento genérico de ordem *l* dado por:

$$\sum_{i=1}^{p}\left[\sum_{i=1}^{n}\frac{z_{ij}z_{il}}{Var(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right]\beta_j^{(k)} + \sum_{i=1}^{n}\frac{(y_{i-}\mu_i)z_{ij}}{Var(Y_i)}\frac{\partial\mu_i}{\partial\eta_i}$$

e, portanto, na forma matricial tem-se

$$\Gamma\left(\hat{\beta}^{(k)}\right)\hat{\beta}^{(k+1)} = Z^T W^{(k)}\boldsymbol{u}^{(k)}$$

onde $u^{(k)}$é um vetor com elemento genérico

$$u_i{}^{(k)} = \sum_{i=1}^{n} z_{ij}\,\beta_j^{(k)} + \left(y_{i-}\mu_i{}^{(k)}\right)\frac{\partial\eta_i{}^{(k)}}{\partial\eta_i{}^{(k)}} =$$

$$= \eta_i{}^{(k)}\left(y_{i-}\mu_i{}^{(k)}\right)\frac{\partial\eta_i{}^{(k)}}{\partial\eta_i{}^{(k)}}$$

e a matriz $W^{(k)}$ representa a matriz W definida anteriormente e calculada em $\hat{\mu}^{(k)}$.

Assim, atendendo a que $\Gamma'(\beta) = Z^T W Z$, tem-se a expressão final para a estimativa de β na *(k +1)-ésima* iteração

$$\hat{\beta}^{(k+1)} = (Z^T W^{(k)} Z)^{-1} Z^T W^{(k)} \boldsymbol{u}^{(k)}$$

Note-se, ainda por análise da expressão anterior, que apesar de o elemento genérico de *W* conter ϕ, ele não entra no cálculo de $\hat{\beta}^{(k+1)}$ e portanto pode-se fazer, sem perda de generalidade, ϕ = *1*, quando se está a calcular as estimativas de β. Assim, é irrelevante, para o cálculo de $\hat{\beta}$ o conhecimento ou não do parâmetro de dispersão.

De uma forma resumida o cálculo das estimativas de máxima verosimilhança de β processa-se, iterativamente, em duas etapas:

1. Dado $\hat{\beta}^{(k)}$ (com *k* a iniciar-se em *0*), calcula-se $\boldsymbol{u}^{(k)}$ e $W^{(k)}$.

2. A nova iterada $\hat{\beta}^{(k+1)}$ é calculada.

As iterações param quando é atingido um critério adequado, por exemplo, $\frac{\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|}{\|\hat{\beta}^{(k)}\|} < \varepsilon$ para algum valor de $\varepsilon > 0$ previamente definido. (Turkman & Silva, 2000)

2.3.2 **Odds Ratio**

Dá-se o nome de *odds* de um acontecimento ao quociente entre a probabilidade de sucesso desse acontecimento, π_i, e a probabilidade de insucesso, isto é: $\frac{\pi_i}{1-\pi_i}$. O *odds*, ao contrário da probabilidade, pode tomar qualquer valor positivo, sem limite superior.

O *odds ratio* é definido para dois conjuntos $(x = 0)$ e $(x = 1)$ de dados binários pelo quociente $OR = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$. Também esta medida pode tomar qualquer valor positivo. Se considerarmos o logaritmo temos $log\ OR = log\left(\frac{\pi_1}{1-\pi_1}\right) - log\ (\frac{\pi_0}{1-\pi_0})$ que não é mais do que a diferença dos *logit* nos dois conjuntos.

Para o modelo de regressão logística com a variável independente dicotómica codificada como *1* e *0*, a relação entre o *odds ratio* e o parâmetro de regressão é dada por:

$$OR = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)/\left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

(Hosmer & Lemeshow, 2000)

2.3.3 Teste de hipóteses e seleção e validação de modelos

A abordagem tradicional para o critério de seleção do modelo estatístico consiste em procurar o modelo mais parcimonioso, ou seja, que envolva o mínimo de parâmetros possíveis a serem estimados e que explique a variabilidade da variável resposta. (Hosmer & Lemeshow, 2000) Sendo assim, existe a necessidade de estabelecer estratégias para selecionar o melhor modelo. As mais utilizadas são baseadas na função de verosimilhança, em particular, o Critério de Informação de Akaike (*AIC*) e o Teste da Razão de Verosimilhança (*TRV*).

2.3.3.1 AIC: Critério de Informação de Akaike

O *AIC* é baseado na função *log-verosimilhança*, com introdução de um fator de correção como modo de penalização da complexidade do modelo. A estatística correspondente é definida por

$$AIC = -2[log(L) - k]$$

onde k é o número de parâmetros do modelo, e L toma o valor da verosimilhança para o modelo estimado.

Quando comparando dois modelos, aquele que apresentar menor valor de *AIC*, ou seja, menos informação perdida será aquele que estará melhor ajustado.

2.3.3.2 TRV: Teste da Razão de Verosimilhança

O *TRV* é apropriado para testar dois modelos desde que estes estejam encaixados, isto é, todas as variáveis explicativas p de um modelo M_1 estão presentes nas q variáveis explicativas de um modelo M_2, onde $p < q = p + k$, com k a representar a diferença do número de parâmetros. A estatística de teste utilizada é

$$LR = -2\,log(M_1) - (-2\,log(M_2)) = -2\,log(\frac{M_1}{M_2})$$

aproximadamente uma distribuição qui-quadrado em amostras grandes, com k graus de liberdade correspondente à diferença de parâmetros dos dois modelos.

Suponhamos que o modelo M_1 com p parâmetros seja o modelo completo, isto é, o modelo maior. E o modelo mais pequeno, M_2 com q parâmetros é designado por modelo reduzido. Isto significa que este último modelo pode ser obtido igualando alguns parâmetros a zero do modelo completo. Este conjunto de parâmetros igualados a zero especifica a hipótese nula a ser testada.

A região de rejeição é dada por: $LR > \chi^2_{\alpha,k}$.

(Kleinbaum & Klein, 2010)

2.3.3.3 **Deviance e Estatística de Qui-Quadrado de Pearson**

As medidas de qualidade de ajustamento permitem dar uma comparação geral entre os valores observados e os valores ajustados da variável resposta.

O modelo saturado, M_S contém tantos parâmetros quanto o número de observações disponíveis.

A *deviance* é definida como:

$$Dev(\hat{\beta}) = -2ln\,(\frac{\hat{L}_c}{\hat{L}_S})$$

onde $\hat{\beta}$ denota o conjunto de parâmetros estimados, $\hat{L}_c$ a máxima verosimilhança, ambos para o modelo de regressão que está a ser avaliado, e $\hat{L}_S$ a máxima verosimilhança para o modelo saturado.

Esta medida compara, assim, a verosimilhança do modelo avaliado com a verosimilhança do modelo saturado, que prevê perfeitamente os resultados observados. Quanto mais perto estiverem os valores das duas verosimilhanças, melhor será o ajustamento (menor o valor da deviance).

(Kleinbaum & Klein, 2010)

A estatística de qui-quadrado de Pearson é dada por:

$$X^2 = \sum \frac{(o-e)^2}{e}$$

em que *o* representa as frequências esperadas e *e* representa as frequências estimadas.

Quando X^2 é avaliado em relação às frequências esperadas estimadas (substituindo pela distribuição binomial), a estatística é determinada por

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

que é assimptoticamente equivalente à *deviance*, dada pela seguinte definição:

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]$$

(Dobson, 2002)

2.3.3.4 **Teste de Wald**

O teste de Wald é usualmente utilizado quando só existe um parâmetro a ser testado. Suponhamos que queremos testar a hipótese:

$$H_0: \beta_j = 0 \; vs \; H_1: \beta_j \neq 0 \text{ , para algum } j.$$

A estatística de teste baseia-se na divisão do coeficiente de interesse, estimado, pelo erro padrão, isto é,

$$Z = \frac{\widehat{B_j}}{\hat{S}_{\widehat{\beta_j}}} \sim N(0,1)$$

O quadrado desta estatística Z^2 é aproximadamente igual a χ^2, uma chi-quadrado com um grau de liberdade. Assim, a hipótese nula é rejeitada a um nível de significância α, se o valor observado do quadrado da estatística de Wald for superior ao quantil de probabilidade *1*- α, de uma χ_1^2.
(Kleinbaum & Klein, 2010)

2.3.4 **Qualidade do Ajustamento**

É essencial avaliar a qualidade do ajustamento após a seleção do modelo que parece ser o mais adequado.

2.3.4.1 **Sensibilidade e Especificidade**

A sensibilidade e a especificidade são medidas estatísticas que medem a performance da classificação de um teste binário.

A sensibilidade mede a proporção de resultados positivos que foram identificados corretamente.

$$Sensibilidade = \frac{nº\ verdadeiros\ positivos}{nº\ verdadeiros\ positivos + nº\ falsos\ negativos}$$

Por outro lado, a especificidade avalia a proporção de resultados negativos que foram identificados corretamente.

$$Especificidade = \frac{nº\ verdadeiros\ negativos}{nº\ verdadeiros\ negativos + nº\ falsos\ positivos}$$

Estas duas medidas estão inversamente correlacionadas uma com a outra. Um preditor perfeito é descrito como tendo 100% sensibilidade (prevê todos os resultados positivos) e 100% especificidade (prevê todos os resultados negativos). Contudo, é extremamente difícil obter tal preditor. Por isso, o que se faz é maximizar a curva da sensibilidade e a especificidade, de maneira a encontrar o ponto ótimo, mais conhecido como *cut-off*.

2.3.4.2 **Curva ROC**

A cada valor de sensibilidade corresponde um valor de especificidade. Esta relação pode ser ilustrada numa curva *ROC*. A curva *Receiver Operating Characteristic* (*ROC*), ou simplesmente curva *ROC*, corresponde à representação gráfica da sensibilidade ou verdadeiros positivos, versus (1 - especificidade) ou falsos positivos. A área sob a curva *ROC*, denominada por *Area Under Curve (AUC)*, varia entre *0* e *1*, e dá-nos uma medida da capacidade de discriminação do modelo, i.e., distinção entre os sucessos dos insucessos.

Os valores considerados por (Hosmer & Lemeshow, 2000) relativamente ao diagnóstico da capacidade de discriminação de um modelo baseado nos valores de *AUC* são:

- $0.5 \leq AUC < 0.7$ - Modelo sem poder discriminatório
- $0.7 \leq AUC < 0.8$ - Discriminação aceitável
- $0.8 \leq AUC < 0.9$ - Discriminação excelente
- $AUC \geq 0.9$ - Discriminação extraordinária

2.3.4.3 **Resíduos**

Para cada indivíduo, o resíduo equivale à diferença entre os valores observados e os estimados. Se existir uma grande diferença, corresponde a um mau ajustamento.

Para a regressão logística, existem duas formas principais de resíduos correspondentes às medidas de qualidade de ajuste *D* e X^2. Para *m* observações, *m* resíduos podem ser calculados. Seja Y_k o número de sucessos, n_k o número de observações e $\hat{\pi}_k$ a probabilidade estimada de sucesso para a *k* observação, temos que:

$$X_k = \frac{(y_k - n_k\hat{\pi}_k)}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}, \quad k = 1, \dots, m.$$

é o Resíduo de *Pearson*.

Uma vez que $\sum_{k=1}^{m} X_k^2 = X^2$ é a estatística de qualidade de ajustamento de Pearson qui-quadrado, os resíduos de Pearson estandardizados podem ser definidos como:

$$r_{P_k} = \frac{X_k}{\sqrt{1-h_k}}$$

onde h_k são os valores da diagonal da matriz *hat*.

A matriz *hat H,* também designada por matriz de projeções, descreve a influência de cada valor observado tem no valor estimado. *H* é uma matriz *nxn* idempotente ($HH = H = H^2$) e simétrica ($H=H^T$).

Os resíduos de *deviance* podem também ser definidos como:

$$d_k = sign(y_k - n_k\hat{\pi}_k)\left\{2\left[y_k\, log\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k) log\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k}\right)\right]\right\}^{\frac{1}{2}}$$

em que o termo $sign(y_k - n_k\hat{\pi}_k)$ assegura que d_k tem o mesmo sinal que X_k.

Como $\sum_{k=1}^{m} d_k^2 = D$ é a *deviance*, os resíduos de *deviance* estandardizados podem ser definidos como:

$$r_{D_k} = \frac{d_k}{\sqrt{1-h_k}}$$

onde h_k são os valores da diagonal da matriz *hat*. (Dobson, 2002)

2.4 Grupos de Controlo

De forma a se poder estudar a eficácia e a utilidade da atividade de comercialização num conjunto de indivíduos, chamado grupo alvo, é definido um outro grupo distinto, intitulado de grupo de controlo. Este grupo terá como diferença do grupo alvo, não ser testada a atividade de comercialização. Desta forma, a sua diferença comportamental (resposta, pedidos e assim por diante) permitirá avaliar o impacto da atividade comercial em estudo.

2.4.1 Metodologia

O tamanho do grupo de controlo é geralmente o equilíbrio entre a precisão requerida para uma amostra de ensaio e a confiança exigida nos resultados. Se forem necessários mais níveis de precisão e confiança, um tamanho maior do grupo de controlo será necessário.

O sucesso de cada atributo em análise é estimado pela proporção de um dado conjunto de elementos com determinadas características.

Se *C* for o número de elementos com um atributo *A* numa população de dimensão *N*, então a sua proporção *P* é definida por $P = \frac{C}{N}$ e $Q = 1 - P$, sendo Q a proporção dos elementos sem o atributo *A* na população. O estimador natural de *P* é, por conseguinte, $p = \frac{c}{n}$, onde *c* é o número de elementos com atributo *A* na amostra e *n* a sua dimensão.

Os momentos de primeira e segunda ordem do estimador *P* são respetivamente:

- $E(p) = P$
- $Var(p) = \frac{N-n}{N-1} \times \sqrt{\frac{PQ}{n}}$.

2.4.2 Dimensão

O problema para encontrar a dimensão da amostra necessária, reduz-se a manter o erro absoluto do estimador dentro de um limite aceitável, designado por tolerância e representado por *d*.

Se admitirmos que os estimadores são assintoticamente normais então temos que:

$$P(|p - P| > d) \leq \alpha \Leftrightarrow P(|p - P| \leq d) > 1 - \alpha \Leftrightarrow$$

aproximando p a Normal, $Z \sim N(0,1)$,

$$\Leftrightarrow P\left(\left|\frac{p-P}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right| \leq \frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right) \geq 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(|Z| \leq \frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right) \geq 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(|Z| \leq \frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right) \geq 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(|Z| \leq \frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right) \geq 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(|Z| \leq \frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right) \geq 1-\alpha \Leftrightarrow$$

$$\Leftrightarrow \frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}} \geq X_{1-\frac{\alpha}{2}} \Leftrightarrow \left(\frac{d}{\sqrt{\frac{N-n}{N-1}\times\sqrt{\frac{PQ}{n}}}}\right)^2 \geq (X_{1-\frac{\alpha}{2}})^2 \Leftrightarrow$$

$$\Leftrightarrow \left(\frac{d}{X_{1-\frac{\alpha}{2}}}\right)^2 \geq \frac{N-n}{N-1}\times\frac{PQ}{N} \Leftrightarrow$$

$$\Leftrightarrow n \geq \frac{N}{\left(\frac{d}{X_{1-\frac{\alpha}{2}}}\right)^2 \times \frac{N-1}{PQ}+1}$$

Sendo assim, *n,* a dimensão da amostra necessária para manter as características da população tem que ser maior que $\frac{N}{\left(\frac{d}{X_{1-\frac{\alpha}{2}}}\right)^2 \times \frac{N-1}{PQ}+1}$.

(Carvalho)

3. Resultados

3.1 Criação de um modelo para uma campanha

Como já foi referido anteriormente o objetivo desta primeira análise é replicar o modelo para uma determinada oferta de um grupo específico de clientes. Este modelo será posteriormente comparado com o modelo adaptativo resultante para a mesma campanha. Para tal, primeiro é necessário recolher uma amostra representativa da população. Devido à elevada quantidade de informação disponível foi necessário seguir vários filtros e critérios, chegando então a uma amostra com um tamanho mais reduzido que continue a ser representativo das características da população em estudo (por se tratar de informação confidencial não é possível revelar a percentagem, sendo no entanto significativa).

3.1.1 **Metodologia**

Em primeiro lugar procedeu-se à extração da informação respetiva dos clientes em estudo, somente nos meses entre fevereiro e abril de 2012 com o objetivo de se obter uma amostra com informação suficiente para se poder fazer um estudo sem perder as características da população que representa.

O critério utilizado para organizar e selecionar as variáveis é baseado na visão de negócio em escolher aquelas que possam influenciar o cliente a aceitar a campanha em estudo.

Cada cliente tem associado um conjunto de informações que o caracterizam. Chegou-se então a um total de 30 variáveis. Estimar o peso de cada um destes fatores, e a sua influência na decisão do cliente em aderir a uma determinada campanha, é um dos objetivos do nosso estudo.

Para se obter uma imagem da amostra, a sua distribuição aproximada e o papel de cada variável na opção do cliente em rejeitar ou aceitar a campanha, recorreu-se numa primeira fase ao cálculo das estatísticas descritivas, representações gráficas, cruzamento com a variável resposta e testes de homogeneidade. Esta análise preliminar permitiu obter uma melhor perceção da qualidade dos dados e traçar um perfil provisório do cliente. Durante o decorrer da análise foi necessário fazer um reajustamento de algumas das variáveis em estudo consoante o tipo de informação que continham.

3.1.2 **Análise detalhada das variáveis**

A informação utilizada para construir o modelo é relativa a 3834 indivíduos. Destes, 14% aceitaram a campanha e os restantes 86% deram uma resposta negativa. Separa-se agora para análise as variáveis em contínuas e discretas.

3.1.2.1 **Análise variáveis discretas**

Para cada variável, realiza-se o teste do Qui-Quadrado com o objetivo de testar a independência entre a variável em estudo e a variável resposta. Ainda, através do teste de Wald, analisa-se a significância da variável ao introduzi-la unicamente no modelo.

Ao longo das seguintes análises, vamos considerar o valor de significância alfa como 10%. Devido ao caráter sensível da informação o detalhe das análises foi retirado.

3.1.2.2 **Análise variáveis contínuas**

Utiliza-se o teste de Wald para analisar a significância de cada variável ao introduzi-la unicamente no modelo. Vamos continuar a considerar o valor de significância alfa como 10%.

A informação detalhada das análises foi retirada por questões de confidencialidade.

3.1.3 **Formulação do modelo**

O método de regressão *stepwise analysis* é um processo semi-automático de construção do modelo onde acrescenta ou elimina sucessivamente variáveis baseado nas estatísticas *t* dos seus coeficientes estimados.

Existem várias metodologias para obter um modelo parcimonioso que se ajuste bem aos dados e que explique a variabilidade da variável resposta. Dentro destas temos: *backward analysis* em que se parte de um modelo com todas as covariáveis possíveis e se vai eliminando as que não são significativas (através de testes de Wald); a *forward analysis* que consiste em partir do modelo mais simples (nulo) e, a passo e passo, ir testando a inclusão das variáveis e a qualidade do modelo resultante.

Por último, existe a metodologia *stepwise analysis*, em que parte do modelo nulo, tal como na *forward analysis,* e verifica, em cada inclusão de uma nova covariável, a importância das covariáveis já presentes no modelo, uma vez que a entrada de uma determinada covariável pode ter como consequência tornar as restantes prescindíveis, havendo por isso a necessidade de avaliar, em cada passo, a sua remoção.

O método que se optou por empregar foi o *stepwise analysis,* inserindo passo a passo no modelo as variáveis e ir avaliando a importância das restantes. Com o uso de outros métodos, estavam a surgir problemas de convergência no modelo, pelo que este método torna-se o mais indicado para a nossa avaliação.

Utiliza-se o valor de significância de 5% e escolhe-se em cada momento entre o modelo em que tem a nova variável inserida e o modelo anterior, ou seja, sem a nova variável, pelo Critério de Informação de Akaike (*AIC*). Isto significa que aquele que tiver um menor valor de *AIC*, será aquele onde se perde menos informação e portanto estará melhor ajustado. O modelo final resultante obteve um valor de *AIC* igual a 2768.9 e uma *deviance* igual a 3039.386.

Na tabela seguinte encontra-se a síntese das características das variáveis resultantes no modelo final.

Variável	Categorias	$\hat{\beta}$	$\hat{S}_{\hat{\beta}}$	Estatística de Wald *(t)*	*P-Value*
Intercept		-2.386	0.270	-8.853	$<10^{-3}$
Variável 1		0.352	0.101	3.486	$<10^{-3}$
Variável 2	B	-0.179	0.162	-1.106	0.269
	C	-0.513	0.156	-3.286	0.001
	D	-0.776	0.185	-4.187	$<10^{-3}$
	E	-0.871	0.216	-4.096	$<10^{-3}$
Variável 3	Sim	0.037	0.241	0.154	0.878
	Outro	1.479	0.362	4.088	$<10^{-3}$
Variável 4	B	-0.690	0.348	-1.984	0.047
	C	0.437	0.325	1.342	0.180
Variável5	B	0.007	0.261	0.028	0.978
	C	0.139	0.282	0.493	0.622
	Outros	0.897	0.184	4.869	$<10^{-3}$
Variável 6		$<10^{-3}$	$<10^{-3}$	-2.062	0.039
Variável 7		$<10^{-3}$	$<10^{-3}$	9.149	$<10^{-3}$
Variável 8		0.008	0.003	2.901	0.004

Variável 9		0.250	0.109	2.293	0.022
Variável 10		0.029	0.012	2.398	0.017

TABELA 1 - RESUMO DAS PRINCIPAIS MEDIDAS DAS VARIÁVEIS SIGNIFICATIVAS DO MODELO.

3.1.4 Validação do modelo

Calculou-se a *curva ROC* para medir o poder preditivo do modelo.

O valor estimado da probabilidade de sucesso a partir do qual se considera que as variáveis explicativas podem explicar a variável resposta designa-se por ponto de corte e caracteriza-se por ser o valor onde a sensibilidade e a especificidade são mais elevadas. Para o modelo estimado, podemos ver na figura seguinte, que a sensibilidade é cerca de 67% e a especificidade cerca de 71%, obtendo um ponto de corte de 0.14. Isto significa que todos os indivíduos com probabilidade estimada pelo modelo (propensão) superior a 0.14 serão considerados positivos do modelo.

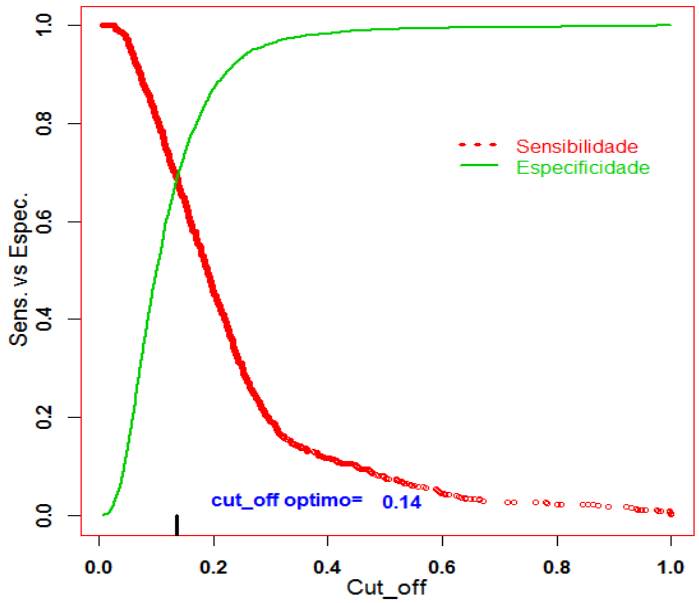

FIGURA 1 - SENSIBILIDADE E ESPECIFICIDADE.

Na figura 26 está representada a *curva ROC*, que cruza a sensibilidade e "1-especificidade", para o modelo estimado. A área sob a *curva ROC*, *AUC*, é aproximadamente 74%, o que corresponde a uma capacidade de discriminação aceitável do modelo.

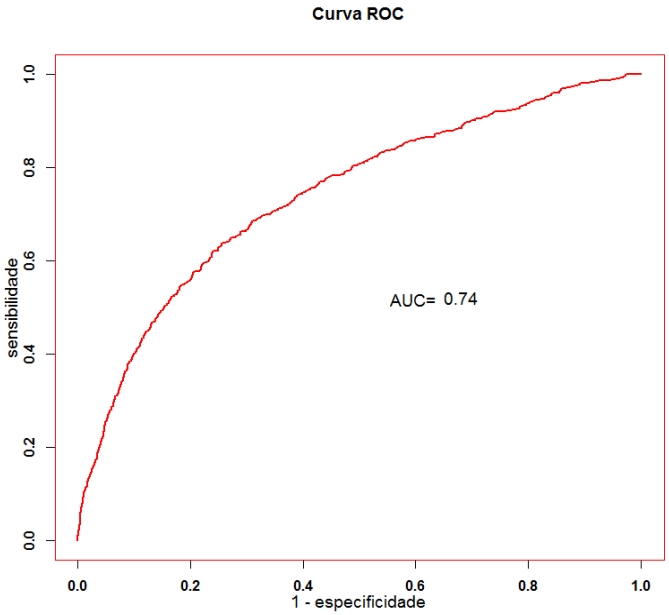

FIGURA 2 - CURVA ROC.

Utilizando os dados em que o modelo foi construído obteve-se para cada observação a propensão correspondente. Tendo em conta o ponto de corte calculado anteriormente como 0.14, comparou-se os indivíduos que aceitaram a campanha versus os indivíduos que são considerados como positivos do modelo, i.e., aqueles com propensão estimada pelo modelo superior a 0.14. A proporção de acertos obtida é então de 69.6%.

$$Proporção\ de\ acertos = \frac{2318 + 351}{3834} = 69.6\%$$

		Observado		
		Rejeita	**Aceita**	**Total**
Estimado	**Propensão < 0.14**	2318	175	**2493**
	Propensão >= 0.14	990	351	**1341**
Total		**3308**	**526**	**3834**

TABELA 2 – COMPARAÇÃO ENTRE OS VALORES AJUSTADOS E ESTIMADOS COM OS DADOS DO MODELO.

Aquando da extração de dados para a construção do modelo, pôs-se de parte um conjunto de indivíduos (1224) para posteriormente se proceder a uma validação do ajustamento do modelo. Assim, para o novo conjunto de dados, com base no modelo final, calculou-se a propensão de cada cliente aderir à campanha e confrontou-se os resultados com os valores observados. Todos os clientes com probabilidade estimada pelo modelo (propensão) superior a 0.14 (ponto de corte calculado anteriormente) foram considerados positivos pelo modelo. Na tabela seguinte podemos verificar as diferenças entre os valores ajustados e os reais. A proporção de acertos para este conjunto de dados foi de 66.6%, um valor semelhante ao obtido de 69.6%, com os dados utilizados pelo modelo.

$$Proporção\ de\ acertos = \frac{698 + 117}{1224} = 66.6\%$$

		Observado		
		Rejeita	**Aceita**	**Total**
Estimado	Propensão < 0.14	698	63	**761**
	Propensão >= 0.14	346	117	**463**
Total		**1044**	**180**	**1224**

TABELA 3 - COMPARAÇÃO ENTRE OS VALORES AJUSTADOS E ESTIMADOS COM O NOVO CONJUNTO DE DADOS.

3.2 Análise das campanhas durante os 3 meses de Verão

Fez-se uma análise a todas as campanhas realizadas entre julho e setembro de 2011 com o intuito de por, um lado, fazer uma síntese do tipo de clientes, campanhas e canal em que estes eram contactados e, por outro, ter uma visão da frequência desses contactos alvos das campanhas. Esta análise foi executada no programa Sql Developer.

A globalidade das campanhas dirigiu-se a clientes A (97.5%) e a clientes B (2.5%).

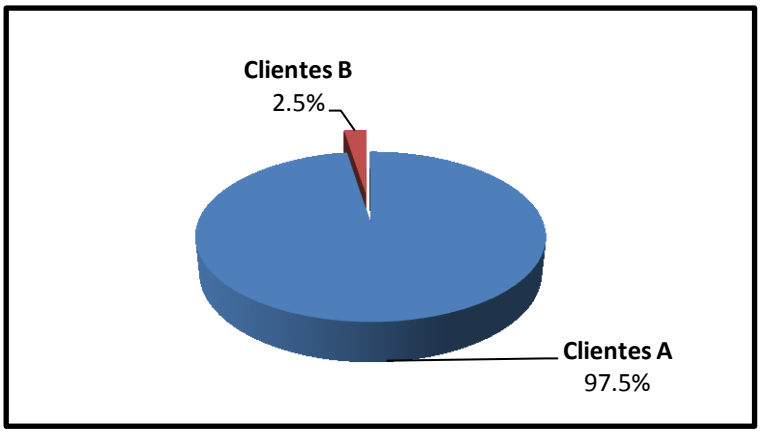

FIGURA 3 - CAMPANHAS POR TIPO DE CLIENTE.

Podemos repartir também as interações por segmentos de clientes a que foram alvos: clientes do segmento 1 (95%), segmento 2 (5%) e outros (0.03%).

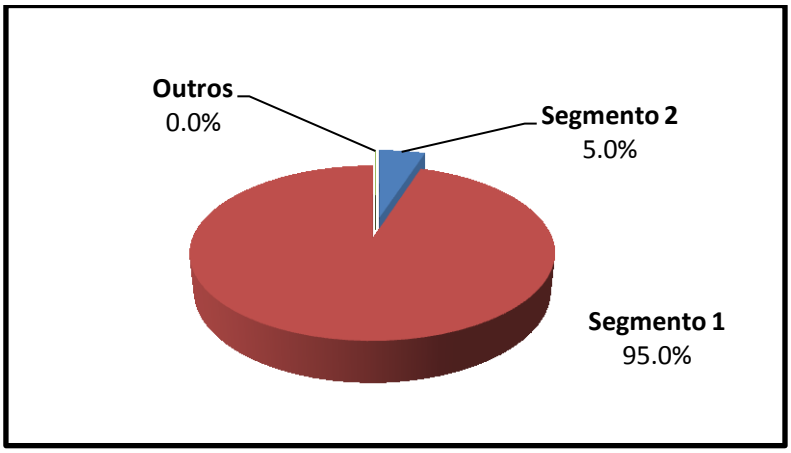

FIGURA 4 - CAMPANHAS POR SEGMENTO DE CLIENTES.

Com o propósito de perceber qual o fim a que se destinavam as campanhas, foi ainda feita uma divisão por tipo de campanha e canal de contacto.

Na primeira, constatou-se que a maioria das campanhas feitas nos meses de Verão tinha um intuito do tipo 1 (94.6%), e as restantes, ou eram campanhas tipo 3 (2.8%), tipo 2 (1.4%) ou de tipo 4 (1.2%).

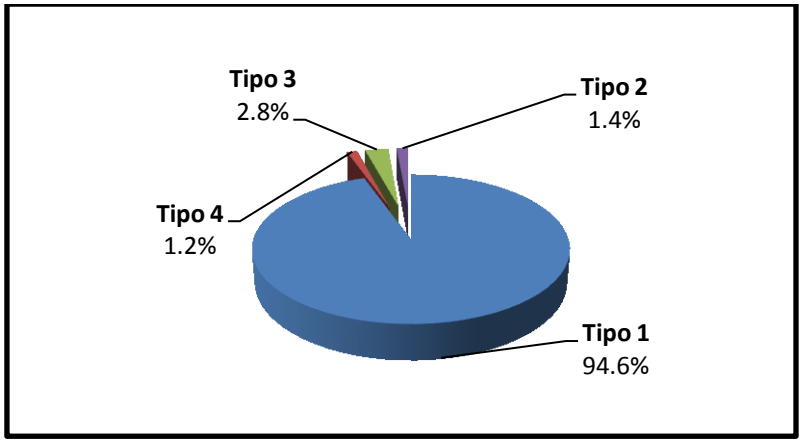

FIGURA 5 - CAMPANHAS POR TIPO.

Em relação ao tipo de canal em que os clientes eram contactados, a generalidade das campanhas foi através de canal 1 (93.1%), 2.6% utilizando o canal 5, 2.5% através do canal 3, 1.3% pelo canal 2 e 0.6% pelo canal 4.

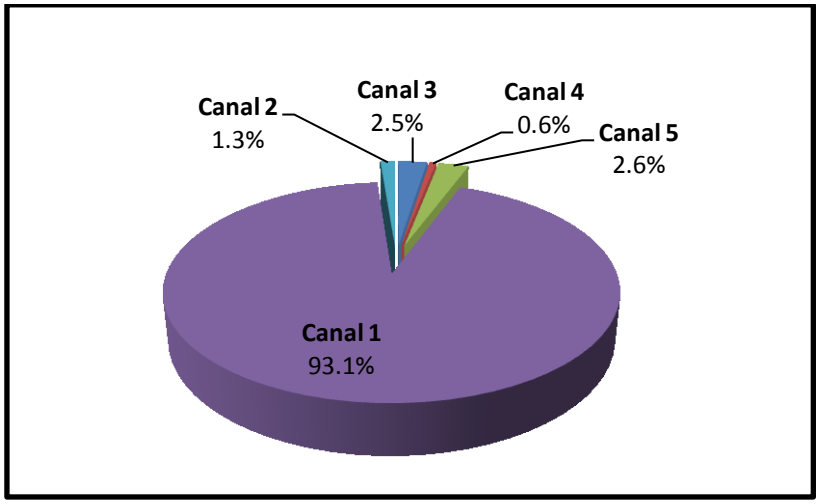

FIGURA 6 - CAMPANHAS POR CANAL.

Clientes A

Passamos agora a fazer uma análise mais detalhada aos clientes A. Dentro dos clientes A e numa análise por segmento, podemos verificar que as campanhas foram feitas na maioria aos clientes de segmento 1 (96.7%), enquanto 3.3% foi feita a clientes segmento 2 e 0.0% aos outros.

Segmento	Percentagem
Clientes Segmento 2	3.3%
Clientes Segmento 1	96.7%
Outros	0.0%

TABELA 4 - CAMPANHAS POR SEGMENTO DE CLIENTES PARA CLIENTES A.

Relativamente ao tarifário do cliente, 59.5% das campanhas realizadas nos meses de Verão foram feitas a clientes de tarifário 3 e tarifário 1, 14.3% a clientes de tarifário 2 e os restantes 26.2% a clientes com outros tarifários.

Tarifário	Percentagem
Outros	26.2%
Tarifário 1	21.0%
Tarifário 2	14.3%
Tarifário 3	38.5%

TABELA 5 - CAMPANHAS POR TARIFÁRIO PARA CLIENTES A.

Numa análise aos clientes A, agora, por segmento de cliente e objetivo da campanha, podemos aferir o seguinte: no segmento 1, 94.6% das campanhas possuem um objetivo do tipo 1, para o segmento outros, 59.0% das campanhas têm um intuito do tipo 3 e 31.6% do tipo 1. No segmento 2, 58.6% das campanhas tiveram fim do tipo 1, seguido de 32.2% com intuito tipo 3.

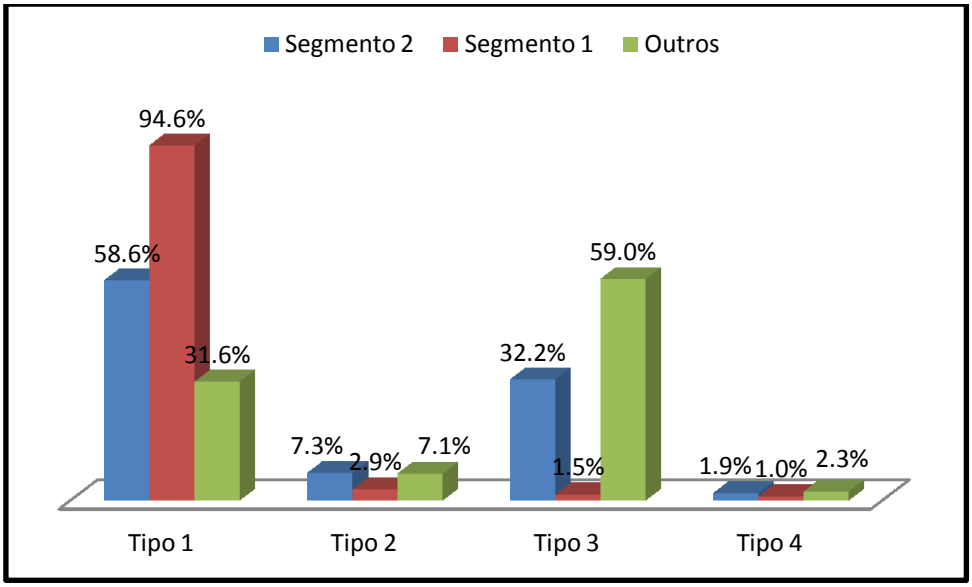

FIGURA 7 - CAMPANHAS POR TIPO EM CLIENTES A.

Por fim, numa análise aos clientes A, por segmento de cliente e canal da campanha, para o segmento 1 95.7% das campanhas foram dirigidas através do canal 1, para clientes de de segmento outros 63.5% das campanhas foram encaminhadas por canal 3 e 26.1% por canal 4. Nos clientes do segmento 2 64.2% das campanhas foram através de canal 1 e 26.5% através de canal 5.

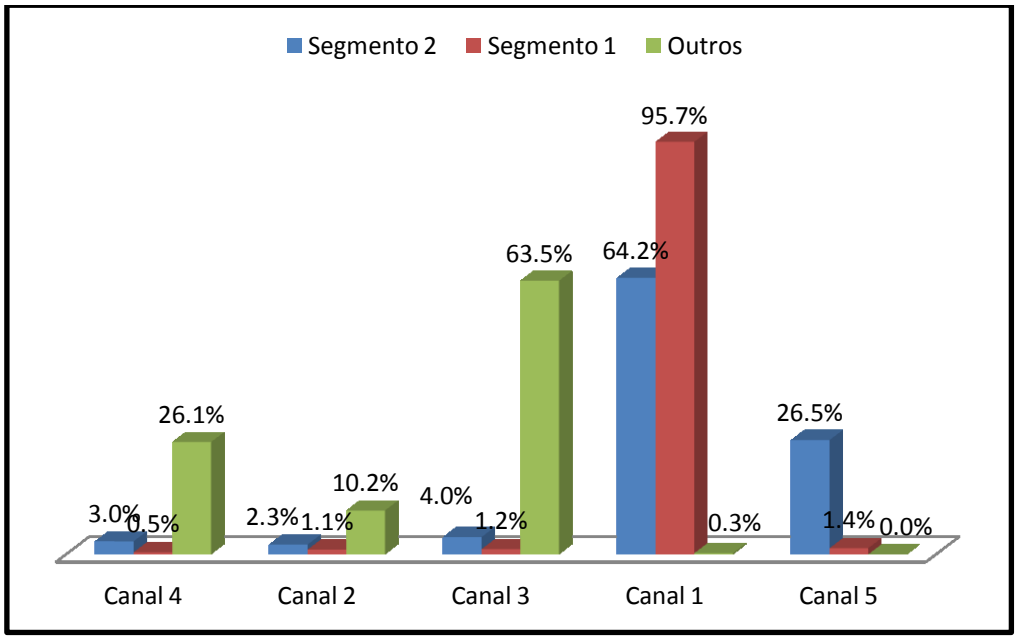

FIGURA 8 - CAMPANHAS POR CANAL EM CLIENTES A.

Em conclusão, verifica-se que os clientes A mantêm as mesmas tendências do total da população em estudo. Por segmento, realiza-se que o segmento 1 é o mais contactado, enquanto, por tarifário, o tarifário 3 é o que regista mais observações. Em relação ao objetivo da campanha, observa-se que em qualquer dos dois segmentos utilizados, o canal 1 têm uma maior percentagem. Por fim, em relação ao canal utilizado, o canal 1 é o que obteve um registo maior para ambos os segmentos.

Clientes B

Examinando em pormenor os clientes B, numa análise por segmento, verifica-se que aproximadamente 72.6% das campanhas foram feitas a clientes do segmento 2, 27.3% a clientes do segmento 1 e 0.1% aos restantes.

Segmento	Percentagem
Clientes Segmento 2	72.6%
Clientes Segmento 1	27.3%
Outros	0.1%

TABELA 6 - CAMPANHAS POR SEGMENTO PARA CLIENTES B.

Relativamente aos tarifários, dentro dos clientes B observou-se que cerca de 55.7% das campanhas foram dirigidas a clientes de tarifário 1 e tarifário 3, 38.3% a outros tarifários e 6.0% a tarifário 2.

Tarifários	Percentagem
Outros	38.27%
Tarifário 1	24.05%
Tarifário 2	6.02%
Tarifário 3	31.66%

TABELA 7 - CAMPANHAS POR TARIFÁRIO PARA CLIENTES B.

Numa análise aos clientes B, agora, por segmento de cliente e objetivo da campanha, nota-se que 53.1% das campanhas dos clientes do segmento 2 foram realizadas com prepósito

do tipo 1 e 22.3% com fim do tipo 2. Em relação aos clientes B do segmento 1, 62.3% das campanhas é feita com o objetivo do tipo l e 19.7% com carácter tipo 3, para clientes de outro tipo, 34.2% das campanhas foram realizadas com o intuito do tipo 2 e 30.4% do tipo 1.

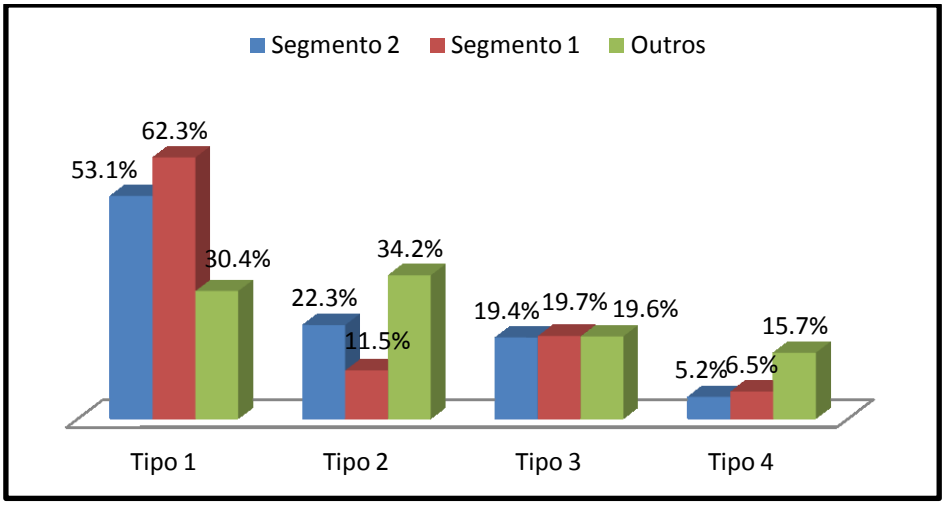

FIGURA 9 - CAMPANHAS POR TIPO PARA CLIENTES B.

Por último, numa análise aos clientes B, por segmento de cliente e canal da campanha, observa-se que 56.7% das campanhas do segmento 1 foram enviadas através do canal 1 e 22.7% pelo canal 3. No segmento outros, 98.6% das campanhas foram remetidas através de canal 3 e em relação aos clientes de segmento 2, 43.7% das campanhas foram feitas através do canal 3, 42.4% por canal 1 e 10.6% através do canal 5.

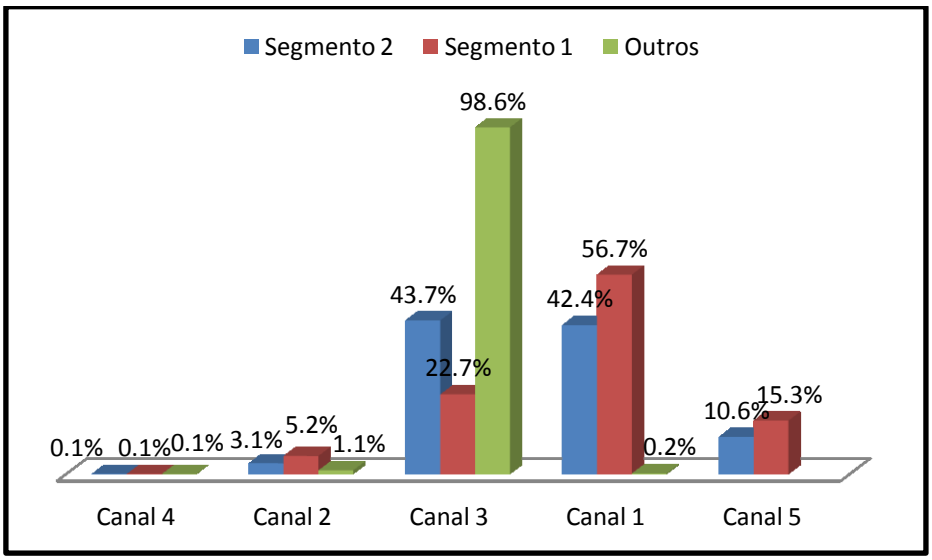

FIGURA 10 - CAMPANHAS POR CANAL PARA CLIENTES B.

Em conclusão, verifica-se que os clientes B não registam as mesmas tendências do total da população em estudo. Por segmento, realiza-se que o segmento 2 é o mais contactado, enquanto, por tarifário, o tarifário 3 é o que regista mais observações. No entanto, em relação ao objetivo da campanha, observa-se que em qualquer dos dois segmentos utilizados, o canal 1 têm uma maior percentagem (comportamento semelhante ao da população). Por fim, em relação ao canal utilizado, o canal 1 obteve um registo maior para o segmento 1, enquanto o canal 3 foi o mais utilizado para o segmento 1.

3.3 Modelos Adaptativos

Os modelos adaptativos são modelos que, consoante as características do cliente, preveem a propensão deste aceitar uma determinada campanha num dado instante. Estes modelos são aplicados em tempo real e ajustados a cada resposta do cliente. Derivam dos modelos *Naive Bayes* e adaptam-se às diferentes interações, mudando as probabilidades ao longo da construção do modelo.

Assim, para cada campanha, existe um modelo adaptativo e define-se, para cada um, uma lista de variáveis que vão ser consideradas na aprendizagem do modelo. Todas as respostas (positivas e negativas) são guardadas juntamente com os detalhes de cada cliente definidos previamente para cada campanha.

Iremos agora abordar as vantagens e desvantagens destes modelos. As vantagens mais relevantes são:

- Algoritmos autossuficientes que conseguem prever sem nenhum histórico de informação;
- Na presença de muitas campanhas, são benéficos para a construção de um modelo preditivo robusto por campanha.

Como desvantagens, temos:

- Modelo inicial errático. Fica estável apenas quando existe informação suficiente para tornar o modelo robusto;
- Os resultados obtidos não são previsões uma vez que o modelo é continuamente adaptável.

O programa de software *Chordiant* estuda estes modelos adaptativos, analisando e explorando os dados, mesmo quando não existe históricos ou quando o comportamento das ofertas é volátil.

O software é uma ferramenta fechada (não há acesso ao código base) e fornece todos os dias 3 relatórios: *Model Performance*, *Predictor Performance Report* e *Active Predictor Report*.

- *Model Performance* dá uma visão geral de todos os modelos. Permite acesso a informações importantes como número total de preditores para cada modelo, preditores ativos e inativos, taxas de sucesso e *CoC* (*Coefficient of Concordance*);

- *Predictor Performance Report* faculta uma medida que avalia para cada variável o seu poder preditivo no modelo;
- *Active Preditor* mostra os preditores ativos e inativos para cada variável de cada campanha, e consequentemente do modelo.

Model Performance

Este *report* é um dos relatórios mais importantes, uma vez que permite ver para cada campanha o número total de preditores de cada modelo, os preditores ativos e inativos, as taxas de sucesso e o *CoC*. A partir destas informações, podemos verificar quais as variáveis que compõem o modelo e o seu poder preditivo e, assim, concluir quais os modelos que necessitam de uma análise mais detalhada de forma a averiguar as causas que penalizam o seu poder preditivo.

MODEL_NAME	AdaptiveModels_C_DD_01	AdaptiveModels_C_DAT_05
PROPOSITION_ID	c_dd_01	c_dat_05
TOTAL_PREDICTORS	39	26
ACTIVE_PREDICTORS	20	15
INACTIVE_PREDICTORS	19	11
LIKELIHOOD	87%	53%
PERFORMANCE_COUNT	1959	1521
PERFORMANCE_POSITIVES	63	645
PERFORMANCE_NEGATIVES	1896	876
SUCCESS_RATE	3%	42%
CoC	66%	69%
CoC_COUNT	1000	1000
CoC_POSITIVES	32	423
CoC_NEGATIVES	968	577

TABELA 8 - EXEMPLO DE REPORT DO *MODEL PERFORMANCE*.

Os parâmetros que compõem o *report* são:

- **MODEL_NAME:** Nome do modelo adaptativo;
- **PROPOSITION_ID:** Abreviatura do modelo adaptativo;
- **TOTAL_PREDICTORS:** Número total de preditores lidos pelo software;
- **ACTIVE_PREDICTORS:** Número de preditores ativos;
- **INACTIVE_PREDICTORS**: Número de preditores inativos;
- **LIKELIHOOD:** Valor da *likelihood* do modelo;

- **PERFORMANCE_COUNT:** Número de casos utilizados para calcular a performance do modelo;
- **PERFORMANCE_POSITIVES:** Número de casos positivos utilizados para calcular a performance do modelo;
- **PERFORMANCE_NEGATIVES:** Número de casos negativos utilizados para calcular a performance do modelo;
- **SUCCESS_RATE:** Indica a taxa de sucesso do modelo (positivos / (positivos + negativos);
- **CoC:** Valor do *CoC* (*Coefficient of concordance*), medida da capacidade de predição do modelo;
- **CoC_COUNT:** Número de casos usados para calcular o *CoC* (1000 por defeito);
- **CoC_POSITIVES:** Número de casos positivos usados para calcular o *CoC*;
- **CoC_NEGATIVES:** Número de casos negativos usados para calcular o *CoC*;

Predictor Performance

Para cada modelo em estudo, o *report Predictor Performance* devolve para cada variável o seu valor preditivo através do *CoC*. Desta forma consegue-se distinguir as variáveis mais significativas (maior poder preditivo) das menos significativas (menor poder preditivo).

Variables/Offers	C_DAT_01	C_DAT_02	C_DAT_03
CUST_CALC_POINTS_AVG	50.00	75.09	50.00
CUST_FLAG_ELECT_INVOICE	50.00	52.08	50.00
CUST_FLAG_LOGIN_MY_VF	50.48	52.08	53.68
CUST_FLAG_SS_EMAIL	50.00	50.00	50.00
SERV_ATRB_PHONE_ANTIQUITY	53.84	54.43	56.63
SERV_ATRB_PHONE_HSDPA_IND	50.41	50.00	50.37
SERV_ATRB_PHONE_MP3_IND	51.01	52.08	52.27
SERV_ATRB_PHONE_RANGE	51.15	52.08	58.03
SERV_ATRB_PHONE_TECH_TYPE	50.36	50.00	50.90
SERV_ATRB_PHONE_TECHNOLOGY	50.38	52.08	58.56
SERV_ATRB_PRICING_PLAN_TYPE	50.00	50.00	50.00
SERV_ATRB_SCORE	53.71	50.00	53.73

TABELA 9 - EXEMPLO DE REPORT DO *PREDICTOR PERFORMANCE*.

Active Predictor

O *report Active Predictor* é de grande importância visto que, para cada modelo, indica quais são as variáveis que estão a ser consideradas para efeitos de predição. As variáveis que no *report* não tiverem valores no output não são consideradas pelo modelo, enquanto as que tiverem 1 significa que estão ativas, i.e., contribuem para o seu poder preditivo. As variáveis com valor 0 estão inativas, ou seja, o modelo não as considera significativas.

Variables/Offers	C_DAT_01	C_DAT_02	C_DAT_03
CLNT_CALC_NUM_AC_SERVICES	1	0	0
CUST_CALC_NUM_ACT_SERV			
CUST_CALC_POINTS_AVG	0	0	0
CUST_FLAG_ELECT_INVOICE			
CUST_FLAG_ELIGIBILITY_IPTV	0	1	0
CUST_FLAG_LOGIN_MY_VF			
CUST_FLAG_SS_EMAIL	0	0	0
SERV_ATRB_PHONE_ANTIQUITY	0	0	0
SERV_ATRB_PHONE_RANGE	0	0	0
SERV_ATRB_PHONE_TECH_TYPE	0	0	0
SERV_ATRB_PHONE_TECHNOLOGY	0	1	1
SERV_ATRB_PRICING_PLAN_TYPE			
SERV_ATRB_SCORE	0	0	0
SERV_ATRB_USER_SEGMENTATION			
SERV_CALC_CONTRACT_TO_END	0	0	0
SERV_CALC_DAYS_WITH_PHONE			

TABELA 10 - EXEMPLO DE REPORT DO ACTIVE PREDICTOR.

Com vista a melhorar e otimizar os modelos adaptativos, procedeu-se à eliminação de variáveis em cada campanha cujo critério de exclusão consistiu em identificar as variáveis sem valor preditivo durante 15 dias. Foram também eliminadas dos modelos as variáveis com performance baixa, as que estão a ser utilizadas por todas as campanhas e as variáveis que não estão presentes em nenhuma.

Por outro lado, algumas variáveis foram acrescentadas aos modelos, consoante o seu significado estar relacionado com o objetivo da campanha. Todos estes procedimentos de otimização dos modelos adaptativos tiveram como base as análises dos *reports* diários (*Model Performance*, *Predictor Performance Report* e *Active Predictor Report*).

Behavior Report

Além dos três *reports* que o programa de software *Chordiant* disponibiliza, podemos ainda ter acesso ao *Behavior Report*. Este *report* inclui a performance do modelo adaptativo e das variáveis preditivas, e descreve a diferença entre os registos positivos e negativos. A análise para cada modelo adaptativo é baseada, por defeito, nos últimos 1000 casos.

No exemplo seguinte relativo à mesma campanha estudada anteriormente (modelo referido em 3.1), registaram-se 87 respostas positivas e 912 negativas, perfazendo um total de 1000 registos. Na coluna *field* observa-se as variáveis que o modelo considera relevantes. A performance é medida através do *Coefficient of Concordance* (*CoC*), que avalia a capacidade do modelo em discriminar os clientes que estão a ser classificados corretamente daqueles que não estão. O *CoC* obtido para a campanha em análise foi de 66.58%.

Sample details

This analysis is based on:

Development cases	1000
positive	87
negative	912

Field Summary

This report documents the profiles of:

Field	Description	Performance
Class Profile		66.58%
CUSTOMER_DAILY.SERV_ATRB_PHONE_ANTIQUITY		57.86%
CUSTOMER_MONTHLY_1.SERV_CALC_P_V_U_O_TMN_6MA		56.92%
CUSTOMER_MONTHLY_1.SERV_CALC_V_CALLS_INB_6MA		55.79%
CUSTOMER_MONTHLY_1.SERV_CALC_V_CALLS_OUT_1		57.72%
CUSTOMER_MONTHLY_1.SERV_CALC_V_CALLS_INB_VDF_6MA		54.85%
CUSTOMER_MONTHLY_1.SERV_CALC_V_UNITS_OUT_6MA		55.89%
CUSTOMER_MONTHLY_2.SERV_CALC_V_UNITS_OUT_VDF_1		54.20%
CUSTOMER_MONTHLY_1.SERV_CALC_ARPU_VOICE_6MA		55.23%
CUSTOMER_MONTHLY_1.SERV_CALC_P_V_U_O_FIX_6MA		56.47%
CUSTOMER_MONTHLY_1.SERV_CALC_ARPU_MESSAGING_6MA		54.37%
CUSTOMER_MONTHLY_1.SERV_CALC_COMMUNITY_SIZE_1		54.53%
CUSTOMER_MONTHLY_1.SERV_CALC_TOTAL_AMT_TOP_UPS_1		53.32%
CUSTOMER_MONTHLY_2.SERV_CALC_V_UNITS_O_VDF_6MA		56.07%
CUSTOMER_MONTHLY_1.SERV_CALC_TOTAL_AMT_TOP_UPS_6M		57.12%
CUSTOMER_MONTHLY_1.SERV_CALC_V_UNITS_OUT_1		55.86%
CUSTOMER_MONTHLY_1.SERV_CALC_SMS_CIRCLE_1		54.43%
CUSTOMER_MONTHLY_1.SERV_CALC_SMS_O_ON_NET_6MA		58.05%
CUSTOMER_MONTHLY_1.SERV_CALC_SMS_MO_1		56.06%
CUSTOMER_MONTHLY_1.SERV_CALC_SMS_OUT_ON_NET_1		54.98%
CUSTOMER_DAILY.SERV_CALC_DAYS_WITH_PHONE		56.73%
CUSTOMER_MONTHLY_1.SERV_CALC_ARPU_MESSAGING_1		54.81%
CUSTOMER_DAILY.SERV_ATRB_PHONE_TECHNOLOGY		57.89%
CUSTOMER_DAILY.SERV_ATRB_PHONE_TECH_TYPE		63.07%
CUSTOMER_DAILY.SERV_ATRB_PHONE_HSDPA_IND		57.82%
CUSTOMER_DAILY.CUST_FLAG_LOGIN_MY_VF		54.49%
CUSTOMER_DAILY.SERV_ATRB_PHONE_RANGE		60.79%
CUSTOMER_MONTHLY_1.SERV_CALC_V_UNITS_OUT_PEAK_1		54.52%

The performance measure is: CoC.

FIGURA 11 – EXEMPLO DO BEHAVIOR REPORT PARA A CAMPANHA EM ESTUDO.

Para cada variável existe um conjunto de medidas estatísticas que analisam as diferenças entre os casos positivos e negativos (*Figura 12*):

Field: CUSTOMER_DAILY.CUST_FLAG_LOGIN_MY_VF
Predictive power: 54.49% CoC

Category	positive		negative		Distribution	Behaviour Analysis				
	Count	Percentage	Count	Percentage	Chart	Behaviour	Z-ratio	0.0	1.0	Lift
N	46	52.87%	566	62.06%		0.08	-1.20			87
Y	41	47.13%	346	37.94%		0.11	1.12			121
Unlisted symbols	0	0%	0	0%		0.00	0.00			0
	87	100%	912	100%		0.09				

FIGURA 12 – EXEMPLO DO BEHAVIOR REPORT PARA UMA VARIÁVEL.

- *Behaviour* é a probabilidade de se obter uma resposta positiva no intervalo considerado;
- O *z-ratio* mede a fiabilidade do comportamento esperado. O *z-ratio* é positivo quando o comportamento esperado é acima da média e negativo quando é esperado abaixo da média;
- *Lift* é o rácio (multiplicado por 100) do comportamento esperado de um certo número de casos num intervalo sobre o comportamento esperado do total dos casos.
- *Distribution Chart* representa a proporção de cada tipo (negativo ou positivo) para cada intervalo. A barra verde indica os casos positivos e a barra vermelha os casos negativos;
- *Behaviour Analysis* é a combinação do *behaviour* com o *z-ratio*. Os pontos representam a probabilidade do comportamento positivo. Os pontos laranjas e vermelhos referem-se a probabilidades significativas e fiáveis.

(Service, Chordiant Adaptive Decision, 2010)

Em suma, podemos concluir que os intervalos e as categorias com maiores pontos vermelhos e laranjas correspondem àqueles que apresentam um comportamento distintivo. Os campos com poucos valores residuais representam aqueles que têm uma relação mais fiável com valor esperado.

3.4 Fórmula de priorização das campanhas

Algumas das campanhas são propostas ao cliente segundo uma fórmula de cálculo que prioriza a ordem pela qual devem ser apresentadas.

A fórmula consiste em:

Likelihood x Marketing Value = Priority Score

Onde:

- *Likelihood* é a propensão de um cliente aceitar a campanha que vem do modelo adaptativo (0 a 1);
- *Marketing Value* é o valor comercial do produto/serviço da campanha
- *Priority Score* é o valor que define a ordem pela qual as campanhas são apresentadas ao cliente.

Como só foi possível ter acesso aos valores de *priority score* e *marketing value*, utilizou-se os valores máximos reais destes para se chegar ao valor da *likelihood* para cada campanha.

Campanha	Marketing Value (V)	Likelihood (P)	Priority Score (P *V)
Campanha1	142.50 €	21%	29.65
Campanha 2	322.24 €	7%	21.7
Campanha 3	39.88 €	24%	9.43
Campanha 4	54.00 €	9%	4.24
Campanha 5	49.50 €	7%	3.64
Campanha 6	4.30 €	51%	2.19
Campanha 7	2.42 €	61%	1.48
Campanha 8	2.50 €	40%	1
Campanha 9	1.06 €	74%	0.78

TABELA 11 - RESULTADOS DE *PRIORITY SCORE* PARA CADA CAMPANHA.

Na análise à tabela anterior, constata-se que as campanhas com *marketing value* elevado têm prioridade em relação às campanhas com *likelihood* mais alta.

Com vista a amenizar este efeito, altera-se a fórmula do cálculo de priorização das campanhas da seguinte forma:

$$(Likelihood \times (\log(Marketing\ Value)+1) = Priority\ Score$$

Onde:

- *Likelihood* é a propensão de um cliente aceitar a campanha que vem do modelo adaptativo (0 a 1);
- *Marketing Value* é o valor comercial do produto/serviço da campanha
- *Priority Score* é o valor que define a ordem pela qual as campanhas são apresentadas ao cliente.

Campanha	Mark. Value (V)	Likelihood (P)	LOG (V) +1	Priority Score (LOG(V)+1)*P
Campanha1	142.50 €	21%	3.15	0.66
Campanha 2	322.24 €	7%	3.51	0.25
Campanha 3	39.88 €	24%	2.60	0.62
Campanha 4	54.00 €	9%	2.73	0.23
Campanha 5	49.50 €	7%	2.69	0.20
Campanha 6	4.30 €	51%	1.63	0.83
Campanha 7	2.42 €	61%	1.38	0.85
Campanha 8	2.50 €	40%	1.40	0.56
Campanha 9	1.06 €	74%	1.03	0.76

TABELA 12 - RESULTADOS DE *PRIORITY SCORE* PARA CADA CAMPANHA COM A FÓRMULA MODIFICADA.

Com esta nova fórmula, as campanhas com *marketing value* elevado continuam a ser prioritárias, contudo o peso da *likelihood* passa a ter influência na ordem pela qual são apresentadas ao cliente.

A função logaritmo (função crescente) permite manter o comportamento dos valores do *marketing value*, minimizando os seus desvios, revelando-se, por isso, uma função adequada para a modelação do *marketing value.* Tal efeito é possível observar-se na figura seguinte.

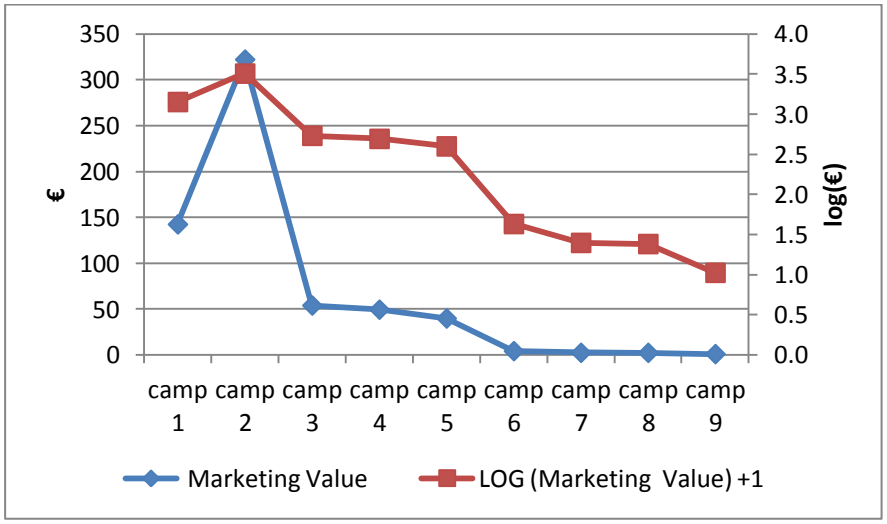

FIGURA 13 – MARKETING VALUE VS LOG(MARKETING VALUE) + 1

3.5 Análise comparativa das taxas de sucesso

De seguida, observa-se a análise comparativa da taxa de sucesso real para as campanhas nas lojas e call center. A taxa de sucesso real foi calculada pelo número de clientes que aceitaram sobre aqueles que foram apresentados. O período em análise foi de 01 de abril de 2011 a 31 de março de 2012.

Na tabela seguinte podemos observar que existem algumas diferenças entre as campanhas nos dois canais.

	Taxa de Sucesso Real (Média Ponderada)	
Categorias	**Lojas**	**Call Center**
Categoria1	32.69%	23.85%
Categoria2	24.47%	N.d.
Categoria3	17.59%	N.d.
Categoria4	16.72%	42.16%
Categoria5	11.30%	n.d
Categoria6	9.95%	25.17%
Categoria7	3.48%	7.33%
Categoria8	2.63%	1.97%
Categoria9	1.90%	1.17%
Categoria10	0.69%	9.95%
Categoria11	0.43%	N.d.
Categoria12	0.23%	N.d.
Categoria13	N.d.	3.39%

TABELA 13 - TAXA DE SUCESSO REAL POR CATEGORIA.

Numa análise por canais verifica-se que o call center apresenta taxas de sucesso acima de 20% para as categorias 1,4,6 enquanto nas lojas apenas as categorias 1 e 2 tiveram uma taxa acima de 20%.

Observando-se por categorias, a 4, 6 e 10 têm uma taxa de sucesso real substancialmente mais elevada no call center que nas lojas. Por outro lado, a categoria 1 é a única que mostra uma taxa muito mais elevada nas lojas que no call center. Por fim, as categorias 7, 8 e 9 registam taxas de sucesso semelhantes entre os dois canais.

As categorias 2, 3, 5,11 e 12 não têm dados no canal de call center devido a estas categorias não serem alvo de campanhas no canal.

3.6 Grupos de Controlo

Um dos objetivos do projeto era a criação de grupos de controlo que visam avaliar a eficiência das campanhas. Ao criar-se um grupo de controlo mais restrito, representando as mesmas características da população, consegue-se avaliar os fatores que levam o cliente a aderir a uma determinada campanha e assim analisar a influência que a campanha tem no cliente alvo.

Não se obteve resultados acerca deste tema, embora na empresa sejam feitos grupos de controlo sempre que necessário e de acordo com as dimensões adequadas, para algumas campanhas. Ainda assim, apresenta-se nesta secção algumas notas de análise e procedimentos para futuros estudos aos grupos de controlo.

A fórmula apresentada no capítulo 2.4 permite calcular o tamanho do grupo de controlo para as campanhas. Para as campanhas onde é conhecida a taxa de sucesso, substitui-se essa taxa por p admitindo um erro de 5% e desta forma, calcula-se o n.

$$n \geq \frac{N}{\left(\frac{d}{X_{1-\frac{\alpha}{2}}}\right)^2 \times \frac{N-1}{PQ} + 1}$$

Se não for possível saber a taxa de sucesso, utiliza-se um grupo de controlo cuja dimensão represente 10% da população alvo. Outra possibilidade é substituir na fórmula anterior a taxa de sucesso pelo valor 0.5 (50%), garantindo assim que a variância seja máxima.

Para cada campanha é ainda preciso definir o conceito de sucesso. Desta forma, consegue-se calcular a taxa de sucesso e assim identificar (dos que foram alvo da campanha) aqueles que aderiram à campanha e os que rejeitaram.

Dentro do grupo de controlo (não foi alvo da campanha) calcula-se os indivíduos que aderiram à campanha, mesmo não tendo sido alvo de contacto promocional. Desta forma, compara-se a taxa de sucesso do grupo de controlo com a taxa do grupo alvo. Se as taxas tiverem valores semelhantes ou for superior no grupo de controlo, significa que a aderência à campanha não é devida ao contato promocional e que se deve reavaliar os objetivos da

campanha e a segmentação do público-alvo. Por outro lado, se a taxa for inferior, então a campanha mostra-se eficaz e deve-se manter no mercado.

4. Conclusões

Com este projeto pretendeu-se chegar à criação de um modelo para uma determinada campanha, metodologias para a avaliação do resultado de campanhas e de grupos de controlo. Estudou-se também os modelos adaptativos e a priorização das campanhas resultantes destes modelos.

A criação do modelo para uma determinada campanha tinha como intuito estudar quais as variáveis que influenciavam a decisão do cliente a aderir à campanha. Do modelo final resultou um total de 10 variáveis significativas.

O modelo ajustou-se de forma razoável aos dados e teve uma capacidade de discriminação aceitável já que se obteve 74% na curva *ROC*. Testou-se ainda a proporção de acertos para um novo conjunto de dados resultando num valor de 66.6%, idêntico ao resultado obtido com os dados utilizados pelo modelo (69.6%).

Para cada campanha existe um modelo adaptativo que estima a propensão do cliente aceitar uma campanha a cada instante, ajustando-se a cada resposta dada. As otimizações feitas nestes modelos tiveram um impacto positivo, levando a um aumento da capacidade preditiva. Para a campanha em estudo, a capacidade preditiva do modelo adaptativo (*CoC*) foi de 66.58%, valor inferior ao valor do poder preditivo obtido no modelo de regressão logística (74%).

Os *reports* resultantes do software *Chordiant* permitem detetar tendências e mudanças de comportamento, como também inspecionar os modelos isoladamente ou comparativamente em termos de poder preditivo. Tal facto permite dizer que é uma ferramenta de grande utilidade para o estudo dos modelos adaptativos.

De um ponto de vista comercial, o interesse da empresa é apresentar, dentro de um leque de campanhas elegíveis, aquela que é mais atrativa ao cliente. Para se chegar a esta, utiliza-se uma fórmula para priorizar as campanhas (*priority score*). Esta estabelece a ordem pela qual as campanhas devem ser oferecidas ao cliente.

Numa breve análise à propensão e ao *priority score* de cada campanha em estudo, constatou-se que as campanhas com *marketing value* elevado tinham prioridade em relação às campanhas com propensão mais alta. De forma a contrapor este fator, alterou-se a fórmula do cálculo de priorização aplicando a função logaritmo ao *marketing value*. Assim, o peso da propensão passa a influenciar de forma mais significativa a ordem pela qual as campanhas são apresentadas. Com base nestes resultados, pode-se concluir que a alteração feita tem um

efeito positivo quer em termos comerciais quer em termos da satisfação dos clientes, uma vez que, ao pesar-se também a propensão, têm-se mais em conta as necessidades do cliente.

Relativamente aos grupos de controlo, um dos objectivos era avaliar a eficiência das campanhas comparando as taxas de sucesso. Como não se obteve resultados concretos acerca deste tema, expôs-se algumas notas de análise e procedimentos na secção 3.6, com vista a futuros estudos. Será ainda pertinente analisar em mais detalhe, no futuro, a interação dos clientes dos grupos de controlo com outras campanhas e aprofundar uma metodologia de grupos de controlo (define-se os grupos de controlo eram únicos para grupos de campanhas semelhantes, se quando um cliente está no grupo de controlo pode receber outras campanhas, entre outros aspetos).

5. Bibliografia

- Bramer, M. (2007). *Principles of Data Mining.* London: Springer.
- Carvalho, L. (s.d.). *Apontamentos da Cadeira de Amostragem.*
- Christensen, R. (1997). *Log-Linear Models and Logistic Regression* (2ªEdição ed.). Springer.
- Cochran, W. G. (1977). *Sampling Techniques.* New York: John Wiley & Sons.
- Dobson, A. J. (2002). *An Introduction to Generalized Linear Models* (2ªEdição ed.). Chapman & Hall/CRC.
- Faraway, J. J. (2006). *Extending Linear Model with R.* Chapman & Hall/CRC.
- Gomes, J. J. (2011). *Apontamentos da Cadeira de Estatística Aplicada.*
- Harell, F. J. (2011). *Regression Modeling Strategies.* New York: Springer.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2ªEdição ed.). New York: USA: A Wiley-Interscience Publication, John Wiley & Sons Inc.
- Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text.* (3ª Edição e.d.) Springer.
- MacKay, D. J. (1992). *Bayesian Methods for Adaptive Models.* California.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2ªEdição ed.). Chapman & Hall/CRC.
- Mitchell, T. (1997). *Machine Learning.* McGraw-Hill .
- Pestana, D. D., & Velosa, S. F. (2008). *Introdução à Probabilidade e à Estatística* (3ª Edição ed.). Lisboa: Fundação Calouste Gulbenkian.
- Rosenbaum, P. R., & Rubin, B. D. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that incorporate the propensity score. *American Statiscal Association*.

- Service, Chordiant Adaptive Decision. (2010). *A new approach to predicting customer behaviors.* PEGA.
- Service, Chordiant Adaptive Decision. (2010). *An Objective Performance Evaluation of Chordiant Predictive Analytics Director.* PEGA.
- Sheskin, D. J. (2000). *Handbook of Parametrical and Nonparametric Statistical Procedures* (Second Edition ed.). Chapman & Hall/CRC.
- Turkman, M. A., & Silva, G. L. (2000). *Modelos Lineares Generalizados - da teoria à práctica.* Lisboa: Edições SPE.
- (s.d.). Obtido de http://docs.oracle.com/cd/E18727_01/doc.121/e13579/T304264T454950.html